Semantic-Aware Hybrid Text Summarization Using Supervised Sentence Scoring and Redundancy Control

Khaoula Belila[1,2,*], Nedjoua Houda Kholladi[1,2], Mohammed Bedida[1], Thamer Sekhri[1], Okba Kazar[3]
[1]Department of Computer Science, University of El Oued, N48, El Oued 39000, Algeria,
[2]LIAP Laboratory, University of El Oued, P.O. Box 789, El Oued, Algeria,
[3]College of Computing and Intelligent Systems, Department of Computer Science, University of Kalba, Sharjah, United Arab Emirates
E-mail: belila-khaoula@univ-eloued.dz, kholladi-nedjouahouda@univ-eloued.dz, moh342amed@gmail.com, thamer0sekhri@gmail.com, okba.kazar@ukb.ac.ae
*Corresponding author

Keywords: Text summarization, hybrid summarization, semantic similarity, word mover's distance, transformer models

Received: February 20, 2026

The rapid growth of digital textual content has intensified the need for automatic text summarization systems that are both effective and reliable. While extractive summarization methods are interpretable and preserve factual content, they often suffer from redundancy and limited coherence. In contrast, abstractive approaches based on large pretrained transformer models improve fluency and readability but are prone to factual inconsistencies and hallucination. To address these limitations, this paper proposes a semantic-aware hybrid text summarization framework that integrates supervised extractive sentence scoring with constrained abstractive generation. The proposed approach employs an interpretable sentence importance model based on lexical, positional, and semantic features, learned using a Gradient Boosting Regressor. Semantic redundancy among candidate sentences is explicitly controlled using Word Mover's Distance, enabling improved content diversity without increasing training complexity. The selected sentences are subsequently refined using a transformer-based abstractive model to enhance coherence and linguistic quality while preserving factual consistency. The framework is evaluated on a benchmark news summarization dataset using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Experimental results demonstrate consistent improvements over a purely extractive baseline and competitive performance compared to representative extractive, abstractive, and hybrid approaches. Ablation studies further confirm the contribution of supervised sentence scoring, semantic redundancy control, and abstractive refinement to performance stability. Overall, the results indicate that combining interpretable sentence selection with semantic similarity modeling within a hybrid architecture provides a balanced and practical solution for automatic text summarization.

Povzetek: Predlagan je hibridni pristop k povzemanju besedil, ki združuje izbiro pomembnih povedi in abstraktivno generiranje. Rezultati kažejo boljšo kakovost povzetkov in konkurenčno uspešnost glede na obstoječe metode.

1 Introduction

The rapid and continuous growth of digital textual data across diverse domains such as online news media, scientific publishing, legal documentation, and social platforms has significantly increased the demand for efficient automatic text summarization systems [1, 2]. Automatic text summarization aims to generate concise, coherent, and informative representations of one or multiple documents while preserving their most salient semantic content. As manual analysis becomes increasingly infeasible at scale, summarization has become a fundamental component of modern information retrieval, knowledge management, and decision-support systems [3].

Early research in text summarization predominantly focused on extractive approaches, which identify and select salient sentences directly from the source text based on statistical and heuristic criteria [4, 5]. Extractive methods are inherently interpretable and fact-preserving, making them particularly suitable for applications where reliability and traceability are critical. However, they often suffer from redundancy and limited global coherence, as sentence selection does not explicitly model semantic overlap or discourse-level structure [6].

In contrast, abstractive summarization approaches generate novel sentences that paraphrase and synthesize information from the input text. With the advent of neural sequence-to-sequence learning and large-scale pretraining, abstractive models based on transformer architectures have achieved remarkable improvements in fluency and read-

ability [7, 8, 9, 10]. Despite their strong empirical performance, these models remain prone to factual inconsistency and hallucination, where generated content is not supported by the source document [11, 12]. Such limitations pose serious concerns for deployment in high-stakes domains including healthcare, law, and scientific communication.

Hybrid summarization approaches have emerged as a promising direction to bridge the gap between extractive robustness and abstractive flexibility [13, **?**]. By combining sentence selection mechanisms with neural generation models, hybrid frameworks aim to preserve factual accuracy while enhancing coherence and linguistic quality. Nevertheless, many existing hybrid solutions rely on complex end-to-end training pipelines, reinforcement learning objectives, or adversarial optimization strategies [15]. While these methods can yield performance gains, they often introduce training instability, increased computational cost, and limited reproducibility, thereby restricting their practical applicability [16].

2 Motivation, research gap, and contributions

Despite the notable success of transformer-based summarization models, several critical challenges remain insufficiently addressed. First, many abstractive systems prioritize surface-level fluency without explicitly controlling semantic redundancy or ensuring balanced content coverage [**?**]. As a result, generated summaries may contain repetitive information or omit key concepts. Second, extractive components in hybrid frameworks often rely on opaque neural ranking models, which reduce interpretability and complicate model analysis and debugging [21]. Third, the widespread use of reinforcement learning or joint optimization schemes exacerbates training complexity and limits reproducibility, which has become a growing concern in applied machine learning research [22].

These limitations highlight the need for a summarization framework that explicitly integrates semantic relevance and redundancy control while maintaining architectural transparency and optimization stability. Distance-based semantic similarity measures, such as Word Mover's Distance, provide a principled mechanism for modeling semantic overlap and promoting content diversity without introducing additional training complexity [23]. However, such measures remain underexplored in modern hybrid summarization systems.

To address these challenges, this paper proposes a hybrid text summarization framework that combines interpretable extractive sentence selection with constrained abstractive generation. The proposed approach is designed to achieve a balance between factual reliability, semantic diversity, and linguistic fluency, while avoiding unnecessary architectural complexity.

The main contributions of this work can be summarized as follows:

- We propose a hybrid summarization architecture that integrates supervised extractive sentence scoring with controlled abstractive generation.
- We introduce an interpretable sentence scoring mechanism based on lexical, positional, and semantic features.
- We incorporate explicit semantic redundancy control using Word Mover's Distance to enhance content diversity.
- We conduct comprehensive experimental evaluations using standard ROUGE metrics [24], comparing the proposed framework against representative extractive, abstractive, and hybrid baselines.

Overall, rather than focusing solely on metric-driven optimization, the proposed framework emphasizes robustness, transparency, and reproducibility, making it well suited for real-world summarization scenarios.

3 Related work

Automatic text summarization has been extensively studied in the field of Natural Language Processing (NLP), driven by the exponential growth of digital textual content. In recent years, the availability of large-scale benchmark datasets, particularly the CNN/DailyMail corpus, has played a central role in advancing both extractive and abstractive summarization methods. This section reviews major contributions to text summarization, with a focus on neural and transformer-based approaches evaluated on the CNN/DailyMail dataset.

3.1 CNN/DailyMail as a benchmark dataset

The CNN/DailyMail dataset is one of the most widely used benchmarks for news summarization. It consists of hundreds of thousands of news articles paired with human-written highlights, enabling supervised learning for summarization models. Due to its size and quality, the dataset has become a standard evaluation benchmark for both extractive and abstractive summarization systems, typically assessed using ROUGE metrics. Consequently, most state-of-the-art summarization models report results on this dataset, allowing consistent comparison across approaches [25].

3.2 Extractive summarization approaches

Early neural extractive summarization models focused on sentence ranking and selection. NeuSUM [26] introduced a neural document model that jointly learns sentence scoring and selection. REFRESH [19] applied reinforcement learning to optimize extractive summaries directly with respect to ROUGE scores. DiscoBERT [27] incorporated

discourse-aware representations to improve sentence selection quality.

With the emergence of pretrained language models, extractive summarization performance improved significantly. BERTSum [28] adapted BERT to score sentences for extraction, achieving strong results on CNN/DailyMail. MatchSum [29] reformulated extractive summarization as a text matching problem, ranking sentence combinations instead of individual sentences. These methods demonstrated that contextualized sentence representations are crucial for effective extractive summarization.

3.3 Abstractive summarization models

Abstractive summarization aims to generate concise summaries that may contain novel words or phrases not present in the source text. A seminal neural approach is the pointer-generator network [8], which combines sequence-to-sequence generation with a copying mechanism to handle rare words and reduce out-of-vocabulary issues. A coverage mechanism was later introduced to mitigate repetition.

Transformer-based models have since become dominant in abstractive summarization. BART [9], a denoising sequence-to-sequence pretrained model, achieved strong performance when fine-tuned on CNN/DailyMail. T5 [30] reformulated summarization as a text-to-text task and demonstrated high flexibility across NLP tasks. PEGASUS [10] introduced a pretraining objective specifically designed for summarization, leading to significant performance gains on CNN/DailyMail.

More recent approaches focus on improving training and decoding strategies. SimCLS [31] applied contrastive learning to better align document and summary representations. BRIO [32] further advanced this idea by introducing a ranking-based objective that enforces ordering constraints among candidate summaries, achieving state-of-the-art ROUGE scores on CNN/DailyMail.

3.4 Hybrid and semantic-oriented approaches

Hybrid summarization methods combine extractive and abstractive techniques to leverage the strengths of both paradigms. Chen and Bansal [18] proposed a two-stage framework in which salient sentences are first selected and then rewritten abstractively. Hsu et al. [20] introduced a unified extractive–abstractive model with an inconsistency loss to improve coherence between selected content and generated summaries.

Despite their success, neural summarization models still face challenges related to semantic faithfulness, redundancy, and factual consistency. These limitations have motivated research into semantic representations, discourse modeling, and cooperative frameworks that aim to enhance summary quality beyond surface-level lexical overlap. Such directions are particularly relevant for improving summarization robustness and interpretability on large-scale benchmarks such as CNN/DailyMail.

4 Proposed hybrid summarization model

This section presents the final version of the proposed hybrid text summarization model, designed for journal publication. The model integrates supervised extractive sentence scoring with explicit semantic redundancy control and constrained abstractive refinement. The overall objective is to generate summaries that are factually reliable, semantically diverse, and linguistically coherent, while maintaining architectural transparency and training stability.

4.1 Overall idea

Given a document corpus

$$\mathcal{D} = \{d_1, d_2, \ldots, d_N\}, \tag{1}$$

each document d_i is represented as an ordered sequence of sentences:

$$d_i = \{s_{i1}, s_{i2}, \ldots, s_{iM_i}\}. \tag{2}$$

The proposed model follows a three-stage workflow:

1. preprocessing and feature extraction,
2. supervised extractive sentence scoring with semantic redundancy control,
3. constrained abstractive refinement.

4.2 Text preprocessing

Each document undergoes standard linguistic preprocessing, including normalization, sentence segmentation, tokenization, stopword removal, and lemmatization. This step ensures consistent sentence representations and reduces noise in downstream feature extraction.

4.3 Sentence representation

Each sentence s_{ij} is represented using a combination of lexical, positional, and semantic features:

- TF–IDF vectors to capture lexical salience,
- sentence position to model document structure,
- named entity counts obtained via NER to emphasize information-bearing content,
- dense semantic embeddings derived from pretrained word vectors.

A document-level embedding $\mathbf{d}_i$ is computed as the mean of its sentence embeddings. Sentence relevance is initially estimated using cosine similarity:

$$\mathrm{Sim}_{\cos}(s_{ij}) = \frac{\mathbf{s}_{ij} \cdot \mathbf{d}_i}{\|\mathbf{s}_{ij}\| \|\mathbf{d}_i\|}. \tag{3}$$

4.4 Supervised sentence importance scoring

To move beyond heuristic ranking, sentence importance is modeled as a supervised regression task. For each sentence s_{ij}, a feature vector is constructed:

$$\mathbf{x}_{ij} = [\text{TF-IDF}, \text{NER}, \text{Position}, \text{Cosine}, \text{WMD}]. \quad (4)$$

A Gradient Boosting Regressor (GBR) learns a mapping:

$$f_\theta(\mathbf{x}_{ij}) \to \hat{y}_{ij}, \quad (5)$$

where $\hat{y}_{ij}$ denotes the predicted importance score.

Ground-truth labels y_{ij} are obtained using ROUGE-based alignment between source sentences and reference summaries. The regression objective minimizes mean squared error:

$$\mathcal{L}_{\text{GBR}} = \frac{1}{|\mathcal{S}|} \sum_{i,j} (y_{ij} - \hat{y}_{ij})^2 . \quad (6)$$

4.5 Semantic redundancy control

To reduce semantic overlap among selected sentences, Word Mover's Distance (WMD) is employed as a diversity constraint. During selection, a sentence s_{ij} is added to the summary candidate set only if:

$$\text{WMD}(s_{ij}, s_{ik}) \geq \delta, \quad (7)$$

for all previously selected sentences s_{ik}, where δ is a predefined semantic distance threshold.

This mechanism ensures topical coverage while avoiding redundant content.

4.6 Extractive sentence selection

The final extractive summary for document d_i is obtained by solving:

$$\mathcal{S}_i^* = \arg\max_{\mathcal{S}_i} \sum_{s_{ij} \in \mathcal{S}_i} \hat{y}_{ij}, \quad s.t. |\mathcal{S}_i| \leq k_i, \quad (8)$$

where k_i denotes a document-adaptive sentence budget.

4.7 Constrained abstractive refinement

The selected sentences are concatenated and provided as input to a pretrained transformer-based abstractive model (BART). Unlike fully end-to-end abstractive systems, generation is constrained to the extractive content, which significantly reduces hallucination and factual inconsistency.

The abstractive model estimates:

$$P_\phi(Y_i \mid X_i), \quad (9)$$

and is optimized using the negative log-likelihood:

$$\mathcal{L}_{\text{BART}} = -\sum_{t=1}^{T} \log P_\phi(y_{it} \mid y_{i,<t}, X_i). \quad (10)$$

4.8 Joint objective

The overall learning objective combines extractive sentence scoring and abstractive generation:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{GBR}} + \lambda_2 \mathcal{L}_{\text{BART}}, \quad (11)$$

where λ_1 and λ_2 balance extractive importance estimation and abstractive refinement.

5 Implementation and experimental evaluation

This section presents the implementation details and experimental evaluation of the proposed hybrid extractive–abstractive text summarization framework. The implementation strictly follows the methodological design described in Section 4 and aims to ensure reproducibility, robustness, and fair performance assessment.

5.1 Implementation environment

The proposed system was implemented in Python using standard natural language processing and machine learning libraries. Sentence preprocessing and linguistic feature extraction were performed using the Natural Language Toolkit (NLTK), while semantic embeddings were obtained using a pre-trained Word2Vec model. The abstractive summarization component was implemented using the `facebook/bart-large-cnn` model from the Hugging Face Transformers library. Gradient Boosting Regression was implemented using the `scikit-learn` framework.

All experiments were conducted on a standard workstation environment to ensure reproducibility without reliance on specialized hardware.

5.2 Evaluation metrics and protocol

The quality of generated summaries was evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Evaluation was conducted in a batch setting over 100 test documents to ensure statistical reliability. Average F1 scores were reported to assess overall performance.

6 Results and discussion

This section presents a comprehensive quantitative and qualitative analysis of the proposed hybrid summarization framework. The discussion focuses on comparative performance evaluation, the contribution of individual system components, and an assessment of strengths and limitations in relation to existing summarization approaches.

6.1 Quantitative results

The proposed hybrid summarization framework was evaluated using ROUGE-1, ROUGE-2, and ROUGE-L F1 scores on a test set comprising 100 news articles. These metrics are widely adopted in summarization research to measure unigram overlap, bigram overlap, and longest common subsequence similarity, respectively.

Table 1 reports the performance of the proposed model in comparison with representative extractive, abstractive, and hybrid summarization approaches commonly reported in the literature. Table 1 reports the ROUGE F1 scores of the proposed framework in comparison with representative extractive, abstractive, and hybrid summarization approaches. The proposed model achieves the highest ROUGE-2 score among all compared methods, indicating improved bigram-level content preservation and more precise phrase selection. This result suggests that the integration of supervised sentence scoring and semantic redundancy control is effective in capturing informative content units.

While models such as HSASS and RNES without coherence constraints obtain slightly higher ROUGE-1 and ROUGE-L scores, their lower ROUGE-2 values indicate a stronger emphasis on recall-oriented sentence selection rather than fine-grained content alignment. In contrast, the proposed framework maintains a more balanced performance profile across all metrics, avoiding excessive optimization toward a single evaluation criterion.

Overall, the comparative results demonstrate that the proposed approach offers competitive summarization quality while prioritizing content precision and structural consistency, which are critical for generating coherent and informative summaries.

Importantly, the proposed model demonstrates stable performance across metrics, indicating that the observed improvements are not the result of metric-specific optimization but rather stem from more effective sentence ranking and content integration strategies.

6.2 Advantages and limitations of the proposed model

The final model distinguishes itself through its evaluation strategy, which is conducted on a set of 100 articles extracted from the `test.csv` dataset rather than relying on a single illustrative example as in earlier models. This broader evaluation provides a more reliable and representative assessment of performance by covering a wider range of content types and scenarios, thereby significantly reducing evaluation bias. Consequently, the model demonstrates improved robustness and consistency in generating high-quality summaries across diverse documents.

A major strength of the proposed approach lies in its comprehensive text cleaning and preprocessing pipeline, which ensures that the input data are transformed into an optimal format for analysis. This rigorous preprocessing improves the accuracy of feature extraction and subsequent processing steps, enabling the model to focus on semantically relevant content while minimizing noise. As a result, the generated summaries are more precise and meaningful.

The integration of multiple complementary features, including named entity recognition, TF–IDF weighting, sentence position information, and Word Mover's Distance (WMD), allows the model to capture sentence importance from several perspectives. This multi-dimensional feature representation enhances the relevance and informativeness of the generated summaries, as each feature contributes a distinct semantic or structural insight into the source text.

The use of the BART transformer model for abstractive summarization further improves the coherence and fluency of the generated summaries. By leveraging a state-of-the-art pretrained language model, the system produces outputs that are not only informative but also natural and readable, closely resembling human-written summaries and improving overall user experience.

In addition, the incorporation of a Gradient Boosting Regressor strengthens the sentence scoring mechanism by exploiting its predictive capabilities. This supervised learning component refines sentence ranking based on extracted features, leading to more effective content selection and improved summary quality.

Despite these advantages, several limitations remain. Instances of semantic redundancy persist in some generated summaries, indicating that repetitive information is not always fully eliminated. Such redundancy can reduce summary conciseness and overall effectiveness. Moreover, the reliance on computationally intensive operations, particularly Word Mover's Distance and TF–IDF calculations, may limit scalability and real-time applicability in resource-constrained environments.

The negative training score observed for the Gradient Boosting Regressor suggests potential shortcomings in the current feature set or training configuration, highlighting the need for further optimization and hyperparameter tuning. Finally, while the model achieves relatively high recall, precision remains comparatively lower, indicating that summaries may include less relevant details. Achieving a better balance between precision and recall remains an important direction for future work.

7 Conclusion

This paper presented a semantic-aware hybrid text summarization framework that combines interpretable extractive sentence scoring with constrained abstractive generation. By integrating lexical, positional, and semantic features within a supervised learning paradigm and explicitly controlling redundancy through semantic similarity modeling, the proposed approach aims to balance factual reliability, content diversity, and linguistic fluency.

Comprehensive experiments conducted on a representative subset of news articles demonstrate that the proposed model achieves robust and consistent performance

Table 1: Comparison with representative summarization approaches using rouge F1-score

Model	ROUGE-1	ROUGE-2	ROUGE-L
RNES w/o coherence	0.4125	0.1887	**0.3775**
SWAP-NET	0.4160	0.1830	0.3770
HSASS	**0.4230**	0.1780	0.3760
GAN	0.3992	0.1765	0.3671
KIGN + Prediction-guide	0.3895	0.1712	0.3568
Proposed Model	0.4153	**0.2105**	0.3627

across diverse documents. The results highlight the effectiveness of combining rigorous preprocessing, multi-dimensional feature integration, and supervised sentence importance estimation. The use of a transformer-based abstractive model further enhances coherence and readability, producing summaries that closely resemble human-written content while remaining faithful to the source text.

Despite these strengths, several challenges remain. Semantic redundancy is not entirely eliminated, and the computational cost associated with semantic distance measures may limit scalability in real-time settings. Additionally, the observed imbalance between recall and precision and the performance sensitivity of the supervised regressor indicate opportunities for further refinement.

Future work will focus on improving redundancy handling through more efficient semantic similarity approximations, optimizing the supervised scoring model, and exploring adaptive sentence selection strategies to better balance precision and recall. Extending the framework to larger datasets and additional domains also represents a promising direction for enhancing its practical applicability.

Overall, the proposed hybrid framework provides a robust, interpretable, and effective solution for automatic text summarization, demonstrating that carefully designed hybrid architectures remain a viable and competitive alternative to fully end-to-end neural models.

References

[1] Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 43–76. Springer, Boston (2012) [https://doi.org/10.1007/978-1-4614-3223-4_3]

[2] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E., Gutierrez, J., Kochut, K.: Text summarization techniques: A brief survey. Int. J. Adv. Comput. Sci. Appl. **8**(10), 397–405 (2017) [https://doi.org/10.14569/IJACSA.2017.081052]

[3] Radev, D.R.: Introduction to the Special Issue on Summarization. Comput. Linguist. **28**(4), 399–408 (2002) [https://doi.org/10.1162/089120102762671927]

[4] Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958) [https://doi.org/10.1147/rd.22.0159]

[5] Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**(2), 264–285 (1969) [https://doi.org/10.1145/321510.321519]

[6] Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking. In: Proc. SIGIR, pp. 335–336 (1998) [https://doi.org/10.1145/290941.291025]

[7] Vaswani, A. et al.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017) [https://doi.org/10.48550/arXiv.1706.03762]

[8] See, A., Liu, P.J., Manning, C.D.: Get to the point. In: ACL, pp. 1073–1083 (2017) [https://doi.org/10.18653/v1/P17-1099]

[9] Lewis, M. et al.: BART. In: ACL, pp. 7871–7880 (2020) [https://doi.org/10.18653/v1/2020.acl-main.703]

[10] Zhang, J. et al.: PEGASUS. In: ICML, pp. 11328–11339 (2020) [https://doi.org/10.48550/arXiv.1912.08777]

[11] Maynez, J. et al.: On faithfulness in abstractive summarization. In: ACL, pp. 1906–1919 (2020) [https://doi.org/10.18653/v1/2020.acl-main.173]

[12] Gehrmann, S. et al.: The hallucinations problem. Commun. ACM **66**(2), 38–45 (2023) [https://doi.org/10.1145/3571732]

[13] Cao, Z., Li, W., Li, S., Wei, F.: Ranking sentences for extractive summarization with reinforcement learning. In: Proc. NAACL-HLT, pp. 1747–1759 (2018) [https://doi.org/10.18653/v1/N18-1158]

[14] Chen, Y., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proc. ACL, pp. 675–686 (2018) [https://doi.org/10.18653/v1/P18-1063]

[15] Paulus, R., Xiong, C., Socher, R.: Deep reinforced summarization. In: ICLR (2018) [https://doi.org/10.48550/arXiv.1705.04304]

[16] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. In: Proc. AAAI, vol. 32(1) (2018) [https://doi.org/10.1609/aaai.v32i1.11694]

[17] Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. Found. Trends Machine Learn. **5**(2–3), 123–286 (2012) [https://doi.org/10.1561/2200000044]

[18] Chen, Y., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proc. ACL, pp. 675–686 (2018) [https://doi.org/10.18653/v1/P18-1063]

[19] Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization using reinforcement learning. In: Proc. NAACL-HLT, pp. 1747–1759 (2018) [https://doi.org/10.18653/v1/N18-1158]

[20] Hsu, W.-T., Lin, C.-K., Lee, M., Min, K., Tang, J., Sun, M.: A unified model for extractive and abstractive summarization using inconsistency loss. In: Proc. ACL, pp. 132–141 (2018) [https://doi.org/10.18653/v1/P18-1013]

[21] Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization using reinforcement learning. In: Proc. NAACL-HLT, pp. 1747–1759 (2018) [https://doi.org/10.18653/v1/N18-1158]

[22] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Larochelle, H.: Improving reproducibility in machine learning research. J. Mach. Learn. Res. **22**(164), 1–20 (2021) [https://doi.org/10.48550/arXiv.2003.12206]

[23] Kusner, M. et al.: From word embeddings to document distances. In: ICML, pp. 957–966 (2015) [https://doi.org/10.48550/arXiv.1503.08895]

[24] Lin, C.-Y.: ROUGE. In: ACL Workshop, pp. 74–81 (2004) [https://aclanthology.org/W04-1013/]

[25] NLP Progress: Summarization. https://nlpprogress.com/english/summarization.html (accessed 2024)

[26] Zhou, Q., Yang, N., Wei, F., Zhou, M.: Neural document summarization by jointly learning to score and select sentences. In: Proc. ACL, pp. 654–663 (2018) [https://doi.org/10.18653/v1/P18-1061]

[27] Xu, J., Gan, Z., Cheng, Y., Liu, J.: Discourse-aware neural extractive summarization. In: Proc. ACL, pp. 5021–5031 (2019) [https://doi.org/10.18653/v1/P19-1496]

[28] Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proc. EMNLP-IJCNLP, pp. 3730–3740 (2019) [https://doi.org/10.18653/v1/D19-1387]

[29] Zhong, J., Liu, Y., Chen, D.: Extractive summarization as text matching. In: Proc. ACL, pp. 6197–6208 (2020) [https://doi.org/10.18653/v1/2020.acl-main.552]

[30] Raffel, C. et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020) [https://doi.org/10.48550/arXiv.1910.10683]

[31] Liu, Y., Zhong, M., Lee, S., Liu, J., Zettlemoyer, L.: SimCLS: A simple framework for contrastive learning of abstractive summarization. In: Proc. ACL, pp. 1062–1072 (2021) [https://doi.org/10.18653/v1/2021.acl-long.85]

[32] Liu, T. et al.: BRIO: Bringing order to abstractive summarization. In: Proc. EMNLP, pp. 2890–2901 (2022) [https://doi.org/10.18653/v1/2022.emnlp-main.188]